1 Preface

This document explains how to extend the model described in Qian [2020] to cover the case where the average gene expressions (grouped by the cell class) in the single cell data are random. We also assume that the cell class prior probability π_k is also unknown and random.

2 The variational distribution

The original joint model as described in Qian is:

$$\begin{aligned}\log \mathrm{p}(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \mathbf{z}) = &- \sum_{g,c,k} \zeta_{c,k} \int \mu_{g,k} e^{-\mathrm{D}_c(t)} \gamma_{g,c} \eta_g dt + \\ &+ \sum_{s,c} z_{s,c} \Big[- \mathrm{D}_c(x_s) + \log(\gamma_{g_s,c}) + \log(\eta_{g_s}) + \sum_k \zeta_{c,k} \log(\mu_{g_s,k}) \Big] + \\ &+ \sum_{g,c} \log \mathrm{p}(\gamma_{g,c}) + \sum_g \log \mathrm{p}(\eta_g) + \\ &+ \sum_{c,k} \zeta_{c,k} \log \pi_k \end{aligned} \tag{1}$$

2.1 The prior on $\mu_{g,k}$

To introduce a random $\mu_{g,k}$ we assume that

$$\mu_{g,k} \sim \Gamma\big(m\mathrm{M}_{g,k},\, m\big)$$

As a reminder we note that

$$\overline{\mu}_{g,k} = \mathrm{E}[\mu_{g,k}] = \mathrm{M}_{g,k}$$

and

$$\overline{\log \mu}_{g,k} = \mathrm{E}[\log \mu_{g,k}] = \psi(m\mathrm{M}_{g,k}) - \log(m)$$

2.2 The prior on π_k

Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ be the vector of cell class probabilities. We assume a Dirichlet prior distribution on $\boldsymbol{\pi}$:

$$(\pi_1, \pi_2, \ldots, \pi_K \mid \delta_1, \delta_2, \ldots, \delta_K) \sim \mathrm{D}(\delta_2, \ldots, \delta_K)$$

and

$$\mathrm{p}(\pi_1, \pi_2, \ldots, \pi_K) \propto \prod_{i=1}^{K} \pi_i^{\delta_i - 1}$$

We remind that

$$\overline{\pi}_k = \mathrm{E}[\pi_k] = \frac{\delta_k}{\sum_{i=1}^{K} \delta_i}$$

and

$$\overline{\log \pi_k} = \mathrm{E}[\log \pi_k] = \psi(\delta_k) - \psi\Big(\sum_{i=1}^{K} \delta_i\Big)$$

2.3 Main

The (new) joint model is:

$$\begin{aligned}
\log \mathrm{p}(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\pi}) = & - \sum_{g,c,k} \zeta_{c,k} \mu_{g,k} A_c \gamma_{g,c} \eta_g + \\
& + \sum_{s,c} z_{s,c} \Big[- \mathrm{D}_c(x_s) + \log(\gamma_{g_s,c}) + \log(\eta_{g_s}) + \sum_k \zeta_{c,k} \log(\mu_{g_s,k}) \Big] + \\
& + \sum_{g,c} \log \mathrm{p}(\gamma_{g,c}) + \sum_g \log \mathrm{p}(\eta_g) + \sum_{g,k} \log \mathrm{p}(\mu_{g,k}) \\
& + \sum_{c,k} \zeta_{c,k} \log \pi_k + \sum_k (\delta_k - 1) \log \pi_k \qquad (2)
\end{aligned}$$

where $A_c = \int e^{-\mathrm{D}_c(t)} dt$.

2.3.1 The variational distribution of $\mu_{g,k}$

The variational distribution is:

$$\begin{aligned}
\log \mathrm{q}(\boldsymbol{\mu}) = & \mathrm{E}_{-\mu} \big[\log(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\pi}) \big] = \\
& - \sum_{g,c,k} \overline{\zeta}_{c,k} \mu_{g,k} A_c \overline{\gamma}_{g,c} \overline{\eta}_g + \sum_{s_g,c} \overline{z}_{s_g,c} \Big[\sum_k \overline{\zeta}_{c,k} \log(\mu_{g_s,k}) \Big] + \sum_{g,k} \log \mathrm{p}(\mu_{g,k}) \qquad (3)
\end{aligned}$$

It is worth reminding that

$$z_{s_g,c} = \begin{cases} 1 & \text{if spot } s_g \text{ from gene } g \text{ belongs to cell } c \\ 0 & \text{otherwise} \end{cases}$$

thus adding together all spots from the same gene yields the total gene counts of gene g for any given cell c. We denote this total gene count by $\mathrm{N}_{g,c}$ and:

$$\mathrm{N}_{c,g} = \sum_{s:s_g=g} z_{s,c}$$

The quantity $\overline{z}_{s_g,c}$ is the expectation of the indicator function defined above and is also equivalent to the probability that spot s_g belongs to cell c. We denote the sum of all those probabilities by $\overline{\mathrm{N}}_{c,g}$ and it expresses the expected counts of gene g for cell c. Hence:

$$\overline{\mathrm{N}}_{c,g} = \sum_{s:s_g=g} \overline{z}_{s,c}$$

In equation 3 the summation over spots can now be expressed in terms of a summation over the genes:

$$\begin{aligned}
\log \mathrm{q}(\mu) =& \mathrm{E}_{-\mu}\big[\log(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\pi})\big] = \\
&= -\sum_{g,c,k} \overline{\zeta}_{c,k}\mu_{g,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + \sum_{g,c} \overline{\mathrm{N}}_{c,g}\Big[\sum_k \overline{\zeta}_{c,k}\log(\mu_{g,k})\Big] + \sum_{g,k}\log \mathrm{p}(\mu_{g,k}) = \\
&= -\sum_{g,c,k} \overline{\zeta}_{c,k}\mu_{g,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + \sum_{g,c,k} \overline{\mathrm{N}}_{c,g}\overline{\zeta}_{c,k}\log(\mu_{g,k}) + \sum_{g,k}\log \mathrm{p}(\mu_{g,k}) = \\
&= \sum_{g,k}\Big[\underbrace{-\sum_c \overline{\zeta}_{c,k}\mu_{g,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + \sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g}\log(\mu_{g,k}) + \log \mathrm{p}(\mu_{g,k})}_{\log \mathrm{q}(\mu_{g,k})}\Big]
\end{aligned} \tag{4}$$

If

$$\mu_{g,k} \sim \Gamma\big(m\mathrm{M}_{g,k},\, m\big) \tag{5}$$

then

$$\log \mathrm{p}(\mu_{g,k}) \propto (m\mathrm{M}_{g,k} - 1)\log(\mu_{g,k}) - m\mu_{g,k}$$

and from equation 4 we have:

$$\begin{aligned}
\log \mathrm{q}(\mu_{g,k}) =& -\sum_c \overline{\zeta}_{c,k}\mu_{g,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + \sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g}\log(\mu_{g,k}) + \\
&+ (m\mathrm{M}_{g,k} - 1)\log(\mu_{g,k}) - m\mu_{g,k} = \\
=& \Big[\sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g} + m\mathrm{M}_{g,k} - 1\Big]\log(\mu_{g,k}) - \Big[\sum_c \overline{\zeta}_{c,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + m\Big]\mu_{g,k}
\end{aligned} \tag{6}$$

Hence:

$$\mu_{g,k} \sim \Gamma\Big(\sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g} + m\mathrm{M}_{g,k}, \quad \sum_c \overline{\zeta}_{c,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + m\Big) \tag{7}$$

If the view expressed by the assignment probabilities $\overline{\zeta}_{c,k}$ is very weak and can be neglected then the (posterior) mean of $\mu_{g,k}$ will tend to $\mathrm{M}_{g,k}$:

$$\overline{\mu}_{g,k} \to \mathrm{M}_{g,k}$$

which is the a-priori and known specification for the mean counts of gene g within class k.

If on the other hand, the assignment probabilities reflect a very strong belief then:

$$\overline{\mu}_{g,k} \to \frac{\sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g}}{\sum_c \overline{\zeta}_{c,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g}$$

The numerator involves the product of the expected gene counts of any given gene g in some cell c by the probability that the cell's class is k. When this product is summed over c it gives us what can be thought of as the model-fitted mean gene counts of g within class k (scaled by the denominator)

In general, the mean of the distribution in 7 is

$$\overline{\mu}_{g,k} = \frac{\sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g} + m\mathrm{M}_{g,k}}{\sum_c \overline{\zeta}_{c,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + m} \tag{8}$$

and the mean of its log-transformation:

$$\begin{aligned}\overline{\log \mu_{g,k}} &= \mathrm{E}[\log \mu_{g,k}] \\ &= \psi\Big(\sum_c \overline{\zeta}_{c,k}\overline{\mathrm{N}}_{c,g} + m\mathrm{M}_{g,k}\Big) - \log\Big(\sum_c \overline{\zeta}_{c,k}A_c\overline{\gamma}_{g,c}\overline{\eta}_g + m\Big)\end{aligned} \tag{9}$$

For our studies we assume that the hyperparameter m of the prior specified in equation 5 is equal to one:

$$m = 1$$

In practice that means that the prior distribution of $\mu_{g,k}$ will yield the same mean as the variance. Also if $m < 1$ then the prior variance widens otherwise it will contract.

2.3.2 The variational distribution of π_k

$$\begin{aligned}
\log \mathrm{q}(\boldsymbol{\pi}) =& \mathrm{E}_{-\pi}\big[\log(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\pi})\big] = \\
=& \sum_{c,k} \overline{\zeta}_{c,k} \log \pi_k + \sum_k (\delta_k - 1) \log \pi_k = \\
=& \sum_k \sum_c \overline{\zeta}_{c,k} \log \pi_k + \sum_k (\delta_k - 1) \log \pi_k = \\
=& \sum_k \Big[\sum_c \overline{\zeta}_{c,k} \log \pi_k + (\delta_k - 1) \log \pi_k \Big] = \\
=& \sum_k \Big[\Big(\sum_c \overline{\zeta}_{c,k} + \delta_k - 1 \Big) \log \pi_k \Big] \qquad (10)
\end{aligned}$$

Thus the posterior of $\boldsymbol{\pi}$ follows a Dirichlet distribution:

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \sim \mathrm{D}\Big(\sum_c \overline{\zeta}_{c,k=\text{class 1}} + \delta_1, \ldots, \sum_c \overline{\zeta}_{c,k=\text{class K}} + \delta_K \Big) \quad (11)$$

The posterior mean of the marginal π_k is

$$\overline{\pi}_k = \mathrm{E}[\pi_k] = \frac{\sum_c \overline{\zeta}_{c,k} + \delta_k}{\sum_k \big[\sum_c \overline{\zeta}_{c,k} + \delta_k \big]} \qquad (12)$$

The expectation $\overline{\zeta}_{c,k}$ is the probability that cell c has class k. Hence the sum

$$\sum_c \overline{\zeta}_{c,k}$$

represents the total number of cells with class k. Thus, from equation 12, the more often a cell type appears the greater the weight it will have in the joint probability model described in equation 2, something that sounds reasonable. We note that the posterior mean of the log marginal is:

$$\overline{\log \pi_k} = \mathrm{E}[\log \pi_k] = \psi\Big(\sum_c \overline{\zeta}_{c,k} + \delta_k \Big) - \psi\Big(\sum_k \Big[\sum_c \overline{\zeta}_{c,k} + \delta_k \Big] \Big)$$